1. Analytics

Amazon Athena

Comece a consultar dados instantaneamente. Obtenha resultados em segundos. Pague somente pelas consultas executadas.

O Amazon Athena é um serviço de consultas interativas que facilita a **análise de dados no Amazon S3 usando SQL** padrão. O Athena não precisa de servidor. Portanto, não há infraestrutura para gerenciar e você paga apenas pelas consultas executadas.

O Athena é fácil de usar. Basta apontar para os dados no Amazon S3, definir o schema e iniciar as consultas usando SQL padrão. A maioria dos resultados é entregue em segundos. Com o Athena, não há necessidade de trabalhos complexos de ETL para preparar dados para análise. Isso permite que qualquer pessoa com experiência em SQL analise conjuntos de dados em grande escala com facilidade e rapidez.

O Athena é fornecido já integrado ao AWS Glue Data Catalog, o que permite criar um repositório de metadados unificado em vários serviços, fazer crawling de fontes de dados para descobrir esquemas e preencher o Catalog com definições novas e modificadas de tabelas e partições, além de manter o versionamento do esquema.

O Athena não usa servidor. Você pode consultar rapidamente os dados, sem necessidade de configurar e gerenciar servidores ou armazéns de dados. Basta apontar para os dados no Amazon S3, definir o schema e começar a consultar dados usando o editor de consultas incorporado. O Amazon Athena permite explorar todos os dados do S3, sem necessidade de configurar processos complexos para extração, transformação e carga (ETL) de dados.

O Amazon Athena usa o Presto compatível com ANSI SQL e funciona com diversos formatos de dados padrão, incluindo CSV, JSON, ORC, Avro e Parquet. O Athena é ideal para consultas rápidas e ad-hoc, mas também pode processar análises complexas, incluindo associações grandes, funções de janela e matrizes.

Características

Ajuda na análise de dados totalmente gerenciado

Muito utilizado em análise de dados em *.csv

Serviço de consulta interativa (utiliza sql)

Serverless (paga por consulta)

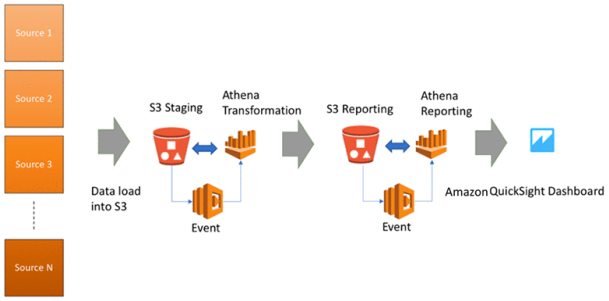
Feito sob o Apache Presto

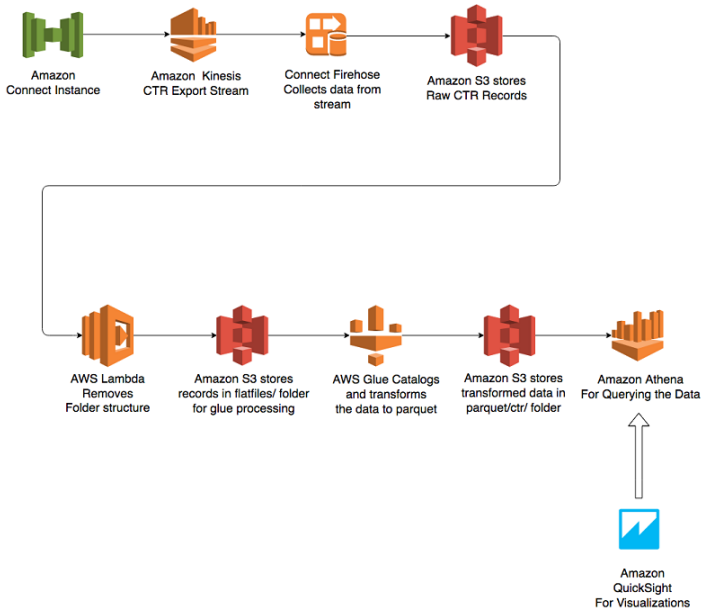
Aponta para o S3

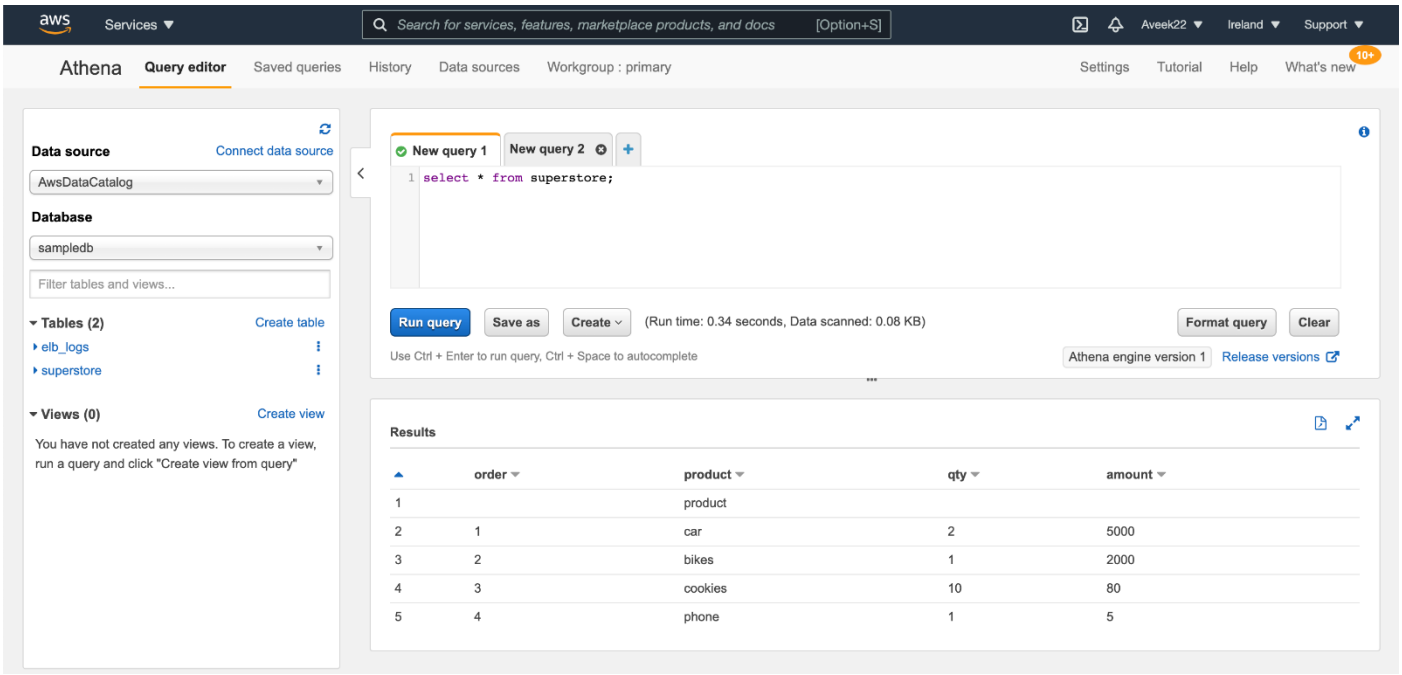
Evita complexidade de ETL

Trocando em miúdos

Basicamente ele cria uma interface com os dados das fontes carregados e o usuário consegue fazer consultas mais facilmente.







➢ Pontos de Atenção

The customer data of an application is stored in an S3 bucket. Your team would like to use Amazon Athena to analyze the data using standard SQL. However, the data in the S3 bucket is encrypted via SSE-KMS. How would you create the table in Athena for the encrypted data in S3? R: Athena decrypts the data automatically, and you do not need to provide key information.

Amazon Elasticsearch Service (Amazon ES)

Elasticsearch é um mecanismo de busca baseado na biblioteca Lucene. Ele fornece um mecanismo de pesquisa de texto completo distribuído com capacidade para vários locatários com uma interface da web HTTP e documentos JSON sem esquema.

O **Elasticsearch** é um mecanismo de busca e análise de dados distribuído, gratuito e aberto para todos os tipos de dados, incluindo textuais, numéricos, geoespaciais, estruturados e não estruturados.

Características Elasticsearch

Base de dados orientada a documentos

Armazenar / buscar / analisar grandes volumes de dados quase que em tempo real

Altamente escalável

Construído com base no Lucene

Open Source e construído em java

Restful

Conceitos Elasticsearch

Cluster: Um grupo de nós (servidores) que guardam dados

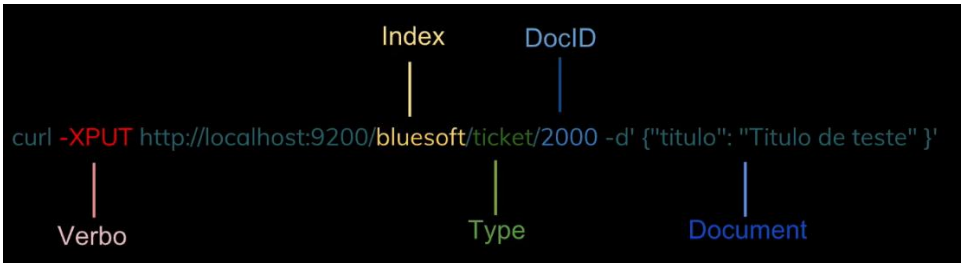
Node: Um servidor individual que armazena informações e faz parte de um cluster

Index: Esqueça o index do SQL. Cada Index do ES index é um agrupamento de documentos

Shards: Um subgrupo de documentos de um index. Um index pode ser dividido em vários shards

Type: É a definição de um schema de um documento dentro de um index (Compare com a tabela no SQL)

Document: Um objeto JSON com dados. É a unidade de informação a ser indexada



Amazon Elasticsearch Service

O Amazon Elasticsearch Service (Amazon ES) é um serviço gerenciado que facilita a implantação, operação e escalonamento de clusters Elasticsearch na nuvem AWS.

Elasticsearch é um mecanismo de pesquisa e análise de código aberto popular para casos de uso como análise de log, monitoramento de aplicativo em tempo real e análise de fluxo de cliques.

Por exemplo, você pode usar o Elasticsearch para adicionar uma caixa de pesquisa ao seu site, analisar logs, métricas e dados de eventos de segurança ou armazenar seus dados para automatizar fluxos de trabalho de negócios.

Com o Amazon ES, você obtém acesso direto às APIs Elasticsearch; o código e os aplicativos existentes funcionam perfeitamente com o serviço.

O Amazon ES provisiona todos os recursos para seu cluster Elasticsearch e o inicia. Ele também detecta e substitui automaticamente nós Elasticsearch com falha, reduzindo a sobrecarga associada a infraestruturas autogerenciadas.

Você pode dimensionar seu cluster com uma única chamada de API ou alguns cliques no console. Para começar a usar o Amazon ES, você cria um domínio. Um domínio Amazon ES é sinônimo de um cluster Elasticsearch.

Os domínios são clusters com as configurações, tipos de instâncias, contagens de instâncias e recursos de armazenamento que você especifica.

Cada instância atua como um nó Elasticsearch. Você pode usar o console Amazon ES para definir e configurar um domínio em minutos.

Amazon EMR

Hadoop é uma plataforma de software em Java de computação distribuída voltada para clusters e processamento de grandes volumes de dados, com atenção a tolerância a falhas. Foi inspirada no MapReduce e no GoogleFS.

Hadoop MapReduce, que é uma ferramenta para processamento e armazenamento de dados massivos.
O **MapReduce** é um modelo de programação que permite o processamento de dados massivos em um algoritmo paralelo e distribuído, geralmente em um cluster de computadores.

MapReduce é um modelo de programação desenhado para processar grandes volumes de dados em paralelo, dividindo o trabalho em um conjunto de tarefas independentes.

MapReduce funciona através de 2 operações: **mapeamento e redução**. No processo de mapeamento (**Map**), os dados são separados em pares (key-value pairs), transformados e filtrados. ... No processo de redução (**Reduce**), os dados são agregados em conjuntos de dados (datasets) menores.

jCaracterísticas

MapReduce é um processo

MapReduce transforma dados

O Job executa programas MapReduce

Amazon EMR (Elastic MapReduce)

Amazon EMR é uma plataforma de cluster gerenciada que simplifica a execução de estruturas de big data, como Apache Hadoop e Apache Spark, na AWS para processar e analisar grandes quantidades de dados.

Usando essas estruturas e projetos de código aberto relacionados, como Apache Hive e Apache Pig, você pode processar dados para fins analíticos e cargas de trabalho de business intelligence.

Além disso, você pode usar o Amazon EMR para transformar e mover grandes quantidades de dados para dentro e para fora de outros armazenamentos de dados e bancos de dados da AWS, como Amazon Simple Storage Service (Amazon S3) e Amazon DynamoDB.

Understanding Clusters and Nodes

O componente central do Amazon EMR é o cluster. Um cluster é uma coleção de instâncias do Amazon Elastic Compute Cloud (Amazon EC2).

Cada instância do cluster é chamada de nó. Cada nó tem uma função dentro do cluster, conhecida como tipo de nó. O Amazon EMR também instala diferentes componentes de software em cada tipo de nó, dando a cada nó uma função em um aplicativo distribuído como o Apache Hadoop.

Master node: Um nó que gerencia o cluster executando componentes de software para coordenar a distribuição de dados e tarefas entre outros nós para processamento. O nó mestre rastreia o status das tarefas e monitora a integridade do cluster. Cada cluster possui um nó mestre e é possível criar um cluster de nó único com apenas o nó mestre.

Core node: Um nó com componentes de software que executam tarefas e armazenam dados no Hadoop Distributed File System (HDFS) em seu cluster. Os clusters de vários nós têm pelo menos um nó principal.

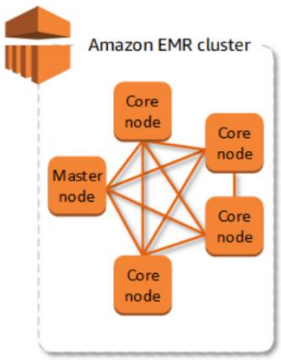
Task node: Um nó com componentes de software que apenas executa tarefas e não armazena dados no HDFS. Os nós de tarefas são opcionais.

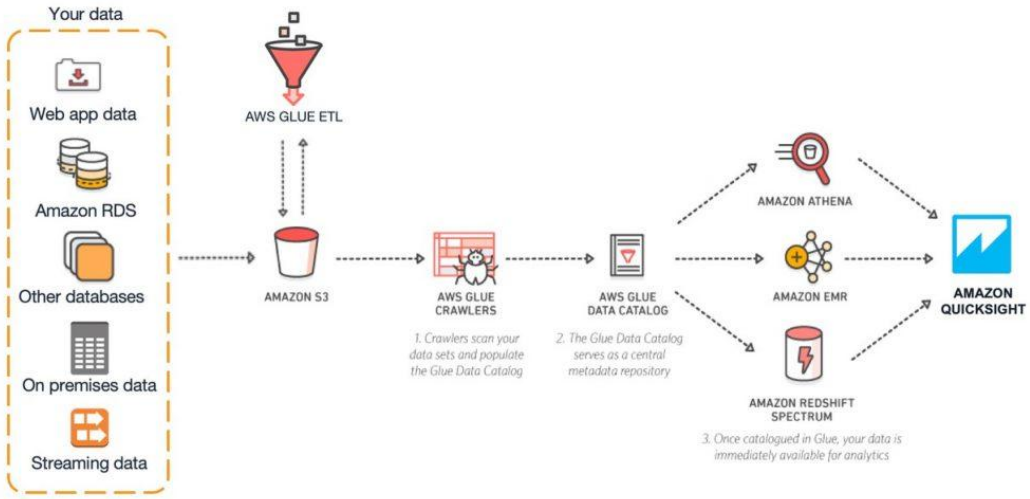
Links Úteis

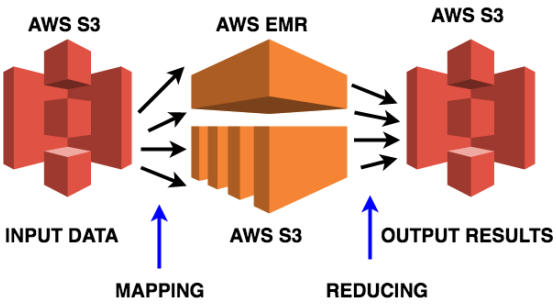
http://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-what-is-emr.html

➢ Pontos de Atenção

You have a set of linux servers running on multiple On-Demand EC2 Instances. The Audit team wants to collect and process the application log files generated from these servers for their report. Which of the following services is the best to use in this case? R: Amazon S3 for storing the application log files and Amazon Elastic MapReduce for processing the log files.







AWS Glue

O AWS Glue percorre suas fontes de dados, identifica os formatos de dados e sugere esquemas para armazenar seus dados. Ele gera o código automaticamente para executar seus processos de transformações e carregamento de dados. Você pode usar o AWS Glue para executar e gerenciar facilmente milhares de trabalhos ETL ou combinar e replicar dados em vários armazenamentos de dados usando SQL.

ETL – Extract, Transform, Load

Extract Transform Load (Extrair Transformar Carregar), são ferramentas de software cuja função é a extração de dados de diversos sistemas, transformação desses dados conforme regras de negócios e por fim o carregamento dos dados geralmente para um Data Mart e/ou Data Warehouse. A extração e carregamento são obrigatórios para o processo, sendo a transformação/limpeza opcional, mas que são boas práticas, tendo em vista que os dados já foram encaminhados para o sistema de destino.

Crawler

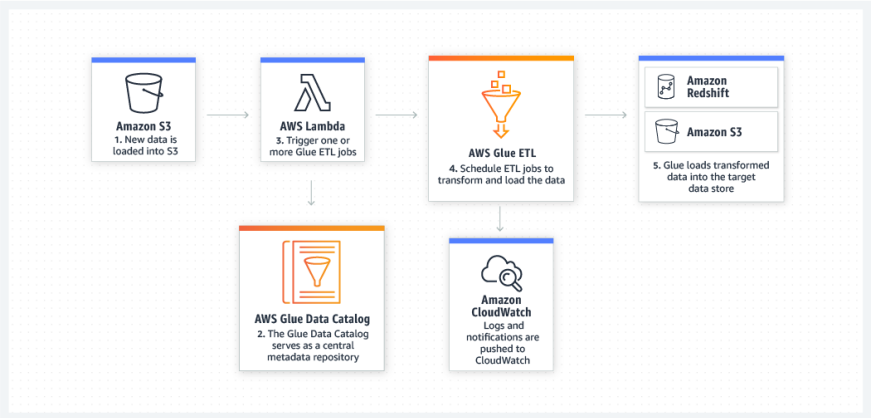
Um rastreador da rede, em inglês web crawler, é um programa de computador que navega pela rede mundial de uma forma metódica e automatizada. Outros termos para rastreadores da rede são indexadores automáticos, robôs, aranhas da rede, robô da rede ou escutador da rede.

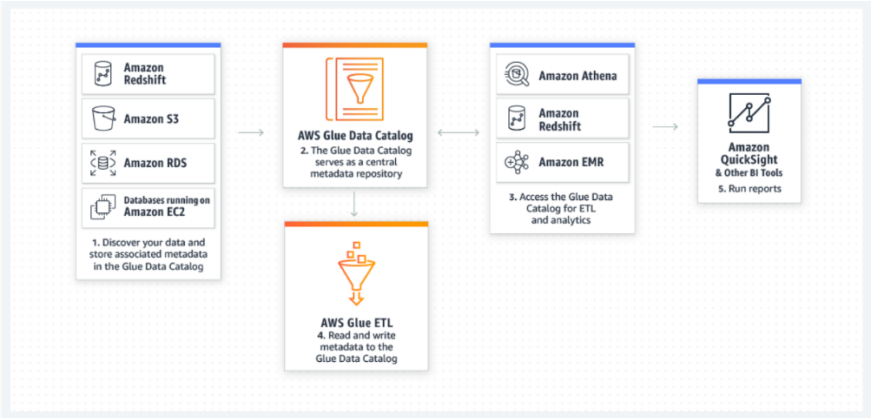
Características

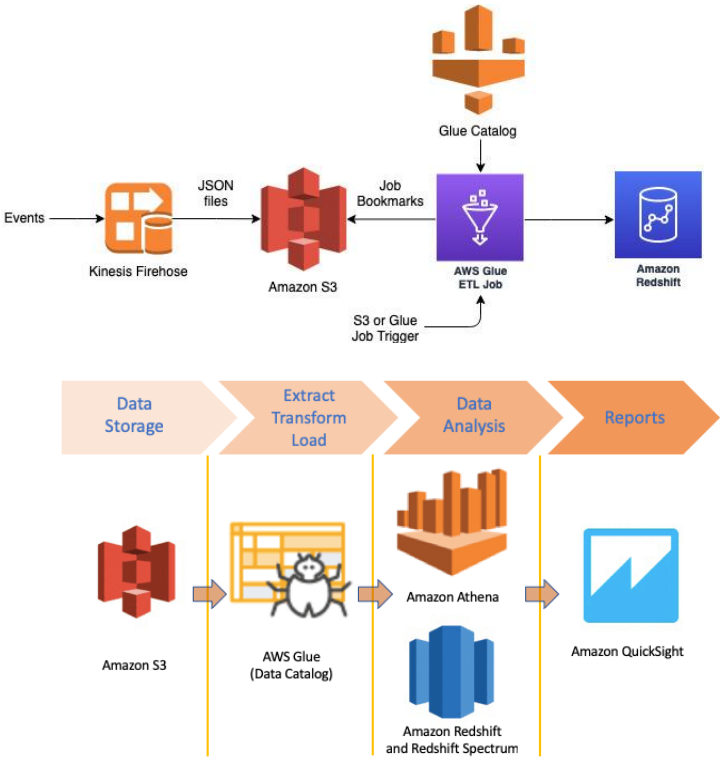
Serviço Serverless gerenciado pela AWS

Serviço de ETL

Possui Crawlers – identifica as origens e gera metadados







➢ Pontos de Atenção

1. Você está trabalhando para uma inicialização de análise de dados que coleta dados de sequência de cliques e os armazena em um depósito S3. Você precisa iniciar uma função AWS Lambda para acionar seus trabalhos ETL para serem executados assim que novos dados estiverem disponíveis no Amazon S3. Qual dos serviços a seguir você pode usar como serviço de extração, transformação e carregamento (ETL) neste cenário? R: AWS Glue

2. Your organization stores customer data in an Amazon DynamoDB table. You need to use AWS Glue to create the ETL (extract, transform, and load) jobs to build the data warehouse. In AWS Glue, you need a service to connect to DynamoDB, determine the schema for the data, and then populate the AWS Glue Data Catalog. Which of the following components should be used to implement it? R: Crawler in AWS Glue. Because, in AWS Glue, you need to add a Crawler that connects to the table and generates the Data Catalog.

Amazon Kinesis

Colete, processe e analise facilmente streams de vídeo e dados em tempo real

O Amazon Kinesis facilita a coleta, o processamento e a análise de dados de streaming em tempo real, permitindo que você obtenha insights oportunos e reaja rapidamente às novas informações.

Com o Amazon Kinesis, você pode consumir dados em tempo real como vídeo, áudio, logs de aplicativos, clickstreams de sites e dados de telemetria de IoT para machine learning, análises e outros aplicativos.

O Amazon Kinesis permite processar e analisar dados assim que são recebidos e responder instantaneamente, em vez de aguardar a conclusão da coleta de dados para poder iniciar o processamento.

Características

Stream: fluxo continuo de dados

Kinesis irá armazenar os dados de streaming

Tipos

Kinesis Stream: Armazena o stream que foram produzidos de 24 h até 7 dias

Kinesis Firehose: Recebe os dados e prontamente disponibiliza aos consumers, descartando logo em seguida

Kinesis Analytics: Analisa os dados que chegam no Stream e no Firehose

Tempo real

O Amazon Kinesis permite consumir, armazenar em buffer e processar dados de streaming em tempo real, proporcionando insights em segundos ou minutos em vez de horas ou dias.

Totalmente gerenciado

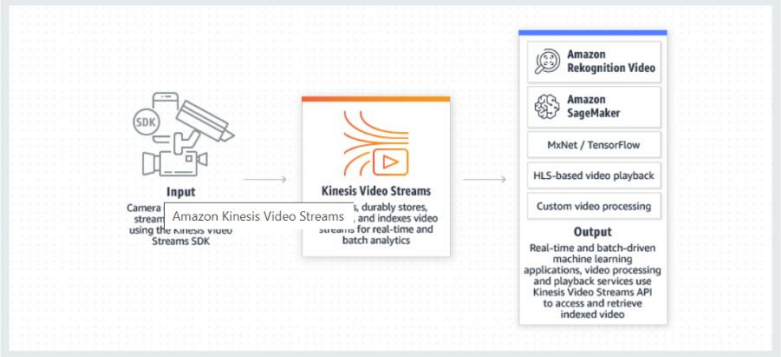
O Amazon Kinesis é totalmente gerenciado e executa aplicativos de streaming sem necessidade de gerenciamento de qualquer infraestrutura.

Escalável

O Amazon Kinesis pode lidar com qualquer quantidade de dados de streaming e processar dados de centenas de milhares de origens com latências muito baixas.

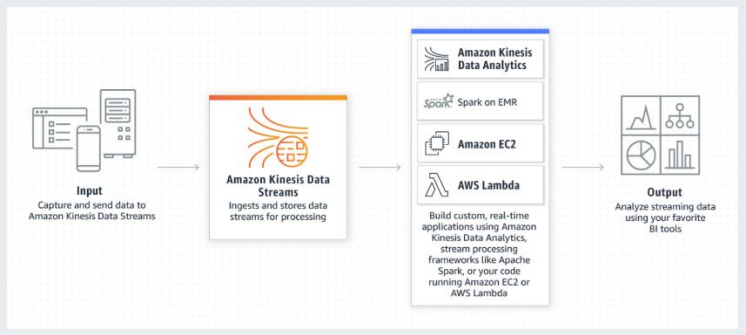
Kinesis Video Streams

Capture, processe e armazene streams de vídeo. O Amazon Kinesis Video Streams facilita o streaming seguro de vídeos de dispositivos conectados para a AWS, onde podem ser usados para análises, machine learning (ML) e outros processamentos.



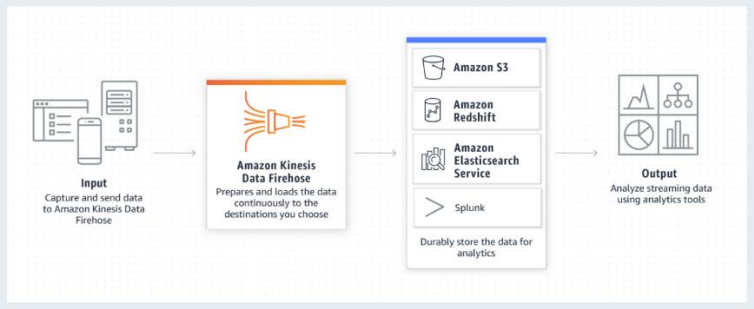
Kinesis Data Streams

Capture, processe e armazene fluxos de dados. O Amazon Kinesis Data Streams é um serviço escalável e durável de streaming de dados em tempo real capaz de capturar continuamente gigabytes de dados por segundo de centenas de milhares de fontes.



Kinesis Data Firehose

Carregue streams de dados em datastores da AWS. O Amazon Kinesis Data Firehose é a maneira mais fácil de capturar, transformar e carregar streams de dados em datastores da AWS para análises praticamente em tempo real usando ferramentas existentes de inteligência de negócios.



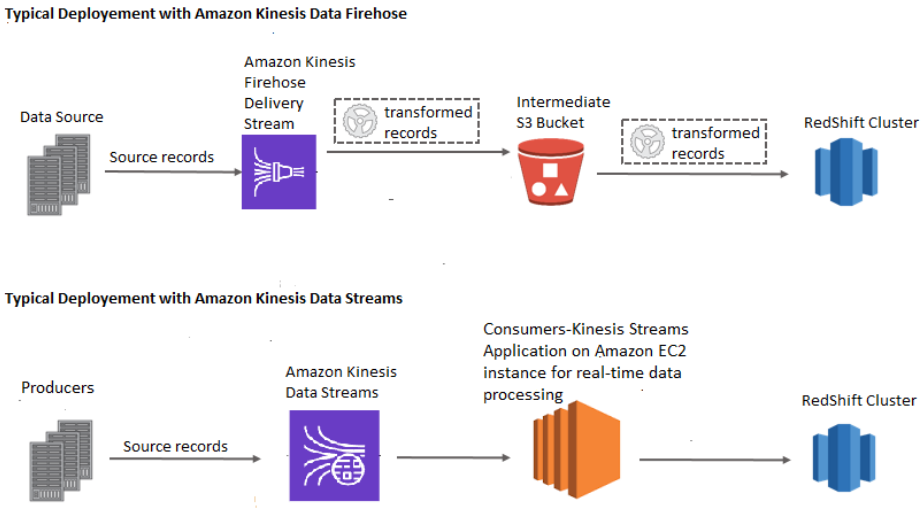
Kinesis Data Analytics

Analise streams de dados com SQL ou Apache Flink. O Amazon Kinesis Data Analytics é a maneira mais fácil de processar streams de dados em tempo real com SQL ou Apache Flink sem a necessidade de aprender novas linguagens de programação ou estruturas de trabalho de processamento.

Links Úteis

https://aws.amazon.com/pt/kinesis/data-streams/faqs/

Data Firehose vs Data Streams



➢ Pontos de Atenção

O Método Heurístico ou de Resolução de Problemas foi criado com o objetivo de encontrar soluções para um problema. Trata-se de simplificar um problema complexo dividindo-o por pequenas partes, de resolução mais fácil, procurando através das respostas de cada uma se possa chegar à resposta ao problema principal

1. Você está trabalhando para uma empresa multinacional de telecomunicações. Seu gerente de TI está disposto a consolidar seus fluxos de log, incluindo os logs de acesso, aplicativo e segurança em um único sistema. Depois de consolidados, a empresa deseja analisar esses logs em tempo real com base em heurísticas. Haverá algum tempo no futuro em que a empresa precisará validar heurísticas, o que requer voltar a amostras de dados extraídas das últimas 12 horas. Qual é a melhor abordagem para atender a esse requisito? R: Primeiro, envie todos os eventos de log para o Amazon Kinesis e, em seguida, desenvolva um processo cliente para aplicar heurísticas nos logs.

2. Você está trabalhando para uma empresa de análise de dados como engenheiro de software, que tem um cliente que está abrindo uma mercearia inovadora sem caixa. Você desenvolveu um aplicativo de monitoramento que usa sensores inteligentes para coletar os itens que seus clientes estão obtendo das geladeiras e prateleiras do supermercado, em seguida, mapeia automaticamente para suas contas. Para saber mais sobre o comportamento de compra de seus clientes, você deseja analisar os itens que são constantemente comprados e armazenar os resultados no S3 para armazenamento durável. Qual serviço você pode usar para capturar, transformar e carregar facilmente dados de streaming no Amazon S3, Amazon Elasticsearch Service e Splunk? R: Amazon Kinesis Data Firehose

3. Você instalou sensores para rastrear o número de visitantes que vão ao parque. Os dados são enviados todos os dias para um fluxo do Amazon Kinesis com configurações padrão para processamento, no qual um consumidor é configurado para processar os dados a cada dois dias. Você notou que seu bucket S3 não está recebendo todos os dados que estão sendo enviados para o stream Kinesis. Você verificou se os sensores estão enviando dados corretamente para o Amazon Kinesis e verificou se os dados são realmente enviados todos os dias. Qual poderia ser o motivo disso? R: Por padrão, os registros de dados são acessíveis apenas por 24 horas a partir do momento em que são adicionados a um fluxo do Kinesis.

4. Você está trabalhando para uma empresa de análise de mídia social como analista de dados chefe. Você deseja coletar gigabytes de dados por segundo de sites e feeds de mídia social para obter insights dos dados gerados por suas ofertas e melhorar continuamente a experiência do usuário. Para atender a esse requisito de design, você desenvolveu um aplicativo hospedado em um grupo Auto Scaling de instâncias Spot EC2 que processa os dados e armazena os resultados no DynamoDB e Redshift. Qual serviço da AWS você pode usar para coletar e processar grandes fluxos de registros de dados em tempo real? R: Amazon Kinesis Data Streams

5. You are planning to launch an application that tracks the GPS coordinates of delivery trucks in your country. The coordinates are transmitted from each delivery truck every five seconds. You need to design an architecture that will enable real-time processing of these coordinates from multiple consumers. The aggregated data will be analyzed in a separate reporting application. Which AWS service should you use for this scenario? R: Amazon Kinesis

6. You are working for a startup that builds Internet of Things (IOT) devices and monitoring application. They are using IOT sensors to monitor all data by using Amazon Kinesis configured with default settings. You then send the data to an Amazon S3 bucket after 2 days. When you checked the data in S3, there are only data for the last day and nothing for the first day. What is the root cause of this issue? R: By default, data records in Kinesis are only accessible for 24 hours from the time they are added to a stream.

7. You are working as a Solutions Architect for a startup in which you are tasked to develop a custom messaging service that will also be used to train their AI for an automatic response feature which they plan to implement in the future. Based on their research and tests, the service can receive up to thousands of messages a day, and all of these data are to be sent to Amazon EMR for further processing. It is crucial that none of the messages will be lost, no duplicates will be produced and that they are processed in EMR in the same order as their arrival. Which of the following options should you implement to meet the startup's requirements? R: Create an Amazon Kinesis Data Stream to collect the messages.

8. Sua empresa tem um aplicativo que foi desenvolvido e precisa ser hospedado em uma instância EC2. A instância EC2 está localizada em uma sub-rede privada e precisa acessar streams AWS Kinesis sem passar para a Internet. Como você pode conseguir isso da melhor maneira possível? R: Create a VPC Interface Endpoint that would allow access to Kinesis Streams

9. Você está trabalhando como arquiteto da AWS para uma empresa financeira global que oferece cotações de negociação de ações em tempo real aos clientes. Você está usando o Kinesis Data Streams para processar feeds do mercado de ações das bolsas de valores e fornecer um painel em tempo real para os clientes. Durante o horário do mercado de ações, muitos usuários estão acessando esses painéis, enquanto no pós-venda, há muitos poucos usuários acessando esses painéis. A equipe de gerenciamento está procurando um número ideal de Kinesis Shards dentro do Kinesis Data Streams. Qual das opções a seguir seria uma solução automatizada para conseguir isso? (Escolha 2) R: 1) Use Application Auto Scaling 2) Use Amazon Kinesis Scaling Utility along with AWS Elastic Beanstalk to automatically modify the number of Shards in Kinesis Data Streams.

   O AWS Application Auto scaling pode ser usado para dimensionar Kinesis Streams automaticamente. Para isso, o CloudWatch pode ser usado para monitorar as métricas de shard do Kinesis Data Stream. Com base nas mudanças nessas métricas, o CloudWatch pode iniciar uma notificação para o Application Auto Scaling. Isso acionará um gateway de API para chamar funções Lambda para aumentar/diminuir o número de fragmentos de fluxo de dados do Kinesis com base em valores métricos. Como alternativa, você pode usar o Amazon Kinesis Scaling Utilities. Para fazer isso, você pode usar cada utilitário manualmente ou automatizado com um ambiente AWS Elastic Beanstalk.

Amazon QuickSight

Serviço BI promovido por ML escalável, sem servidor, construído para a nuvem

O Amazon QuickSight é um serviço de inteligência comercial (BI) promovido por machine learning, escalável, sem servidor, incorporável, construído para a nuvem.

O QuickSight permite que você crie e publique facilmente painéis interativos que incluem o Insights de Machine Learning.

Os painéis do QuickSight podem ser acessados de qualquer dispositivo e incorporados diretamente a aplicativos, portais e sites.

O QuickSight não possui servidor e pode escalar automaticamente para dezenas de milhares de usuários sem nenhuma infraestrutura para gerenciar ou capacidade de planejamento.

Também é o primeiro serviço de BI a oferecer o preço de pagamento por sessão, onde você só paga quando seus usuários acessarem seus painéis ou relatórios, tornando-o custo-eficiente para implantações de grande escala.

Com o QuickSight, você pode fazer perguntas comerciais dos seus dados em linguagem direta e receber as respostas em segundos.